








NDGR: A Noise Divide and Guided Re-labeling Framework for Distantly Supervised Relation Extraction

Zheyu Shi^{1,2}, Ying Mao^{1,2}, Lishun Wang^{1,2}, Hangcheng Li^{1,2},
Yong Zhong^{1,2}, and Xiaolin Qin^{1,2}

¹ Chengdu Institute of Computer Applications, Chinese Academy of Sciences,
Chengdu 610041, China

{shizheyu21,maoying19}@mailsucas.edu.cn, qinxl2001@126.com

² University of Chinese Academy of Sciences, Beijing 100049, China

Abstract. Distant supervision (DS) is widely used in relation extraction to reduce the cost of annotation but suffers from noisy instances. Current approaches typically involve selecting reliable instances from the DS-built dataset for model training. However, these approaches often lead to the inclusion of numerous noisy instances or the disregard of a substantial number of valuable instances. In this paper, we propose NDGR, a novel training framework for sentence-level distantly supervised relation extraction. Initially, NDGR partitions the noisy data from the DS-built dataset by employing a Gaussian Mixture Model (GMM) to model the loss distribution. Afterwards, we utilize a guided label generation strategy to generate high-quality pseudo-labels for noisy data. By iteratively executing the processes of noise division and guided label generation, NDGR helps refine the noisy DS-built dataset and enhance the overall performance. Our method has been extensively evaluated on commonly used benchmarks, and the results demonstrate its substantial improvements in both sentence-level evaluation and noise reduction.

Keywords: Distantly Supervised · Relation Extraction · Label Noise

1 Introduction

Relation extraction (RE) is a fundamental task in the field of Information Extraction (IE), which aims at extracting structured relations between named entity pairs from unstructured text. Most existing methods approach this task by employing supervised training of neural networks, requiring a significant amount of manually labeled data. It is widely acknowledged that data annotation is a

Supported by Sichuan Science and Technology Program (2019ZDZX0006, 2020YFQ0056), Science and Technology Service Network Initiative (KFJ-ST-S-QYZD-2021-21-001), and the Talents by Sichuan provincial Party Committee Organization Department.

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2024
B. Luo et al. (Eds.): ICONIP 2023, CCIS 1969, pp. 98–111, 2024.
https://doi.org/10.1007/978-981-99-8184-7_8

laborious and time-consuming task. In order to address this challenge, Mintz et al. [19] introduced Distant Supervision (DS), an approach that automatically annotates textual data by aligning relation facts extracted from knowledge graphs with the unlabeled corpus. Regrettably, this annotation paradigm inevitably leads to a problem of noise. Hence, there is a need to explore de-noise DSRE methods to minimize the impact of noisy instances.

Currently, there exist two primary approaches for reducing noise in DSRE: the bag-level method and the sentence-level method. The bag-level methods [12, 16, 25, 28, 29] are based on Multi-Instance Learning (MIL), both the training and testing processes are performed at the bag-level. While bag-level approaches are effective in mitigating the influence of noisy data, they do not assign specific labels to each sentence within the bag. Additionally, these approaches overlook cases where all sentences in the bag are false positive samples [23]. These limitations hinder the application of RE in downstream tasks that necessitate sentence-level relation types. Therefore, over the past few years, there has been a growing interest in sentence-level DSRE methods. Most existing sentence-level DSRE methods [4, 8, 11, 22, 31] employ adversarial learning, reinforcement learning, or frequent patterns to filter out noisy data. Although these methods are effective at handling noisy data, they have certain limitations, including reliance on prior knowledge, subjective sample construction, and accessing external data.

In this paper, we propose NDGR, a training framework for sentence-level DSRE. In contrast to previous methods, our method is independent of prior knowledge or external data. It can automatically identify noisy instances and re-label them during training, thereby refining the dataset and enhancing performance. Specifically, NDGR first divides the noisy instances from DS-built data by modeling the loss distribution with a GMM. Since noisy instances contain valuable information, we consider them as unlabeled data and employ a guided label generation strategy to produce high-quality pseudo-labels for the purpose of transforming them into training data. Although the above-mentioned method can improve performance, it fails to fully exploit the potential of each component. Hence, we further design an iterative training algorithm to fully refine the DS-built dataset through iterative execution of noise divide and guided label generation.

The main contributions of this paper are as follows:

- We propose the use of the Gaussian Mixture Model to model the data loss distribution for sentence-level distant supervised RE, which effectively separates noisy data from DS-built data.
- We design a guided label generation strategy, with the aim of generating high-quality pseudo labels to avoid the gradual drift problem in the distribution of sentence features.
- We develop NDGR, a sentence-level DSRE training framework, which combines a noise division, guided label generation, and iterative training to refine DS-built data.

- Our proposed method makes a great improvement over previous sentence-level DSRE methods on widely used datasets, not only relation extraction ability but also noise filtering ability.

2 Related Work

To address the issue of insufficient annotated data in relation extraction, Mintz et al. [19] firstly align unlabeled text corpus with structured data to automatically annotate data. While this method has the capability to autonomously label data, it is bound to engender the issue of wrong labeling. To minimize the impact of mislabeled data, Riedel et al. [25] relaxes the basic assumption of DS to the At-Least-One assumption and applied Multi-Instance Learning [10, 25] to the task of DSRE. In MIL, all sentences with the same entity pair are put into a bag, and assumes that at least one sentence in the bag expresses the relation. Existing bag-level approaches have mainly focused on mitigating the impact of potentially noisy sentences within the bag. Some methods [9, 14, 16, 28] utilize the attention mechanism to assign different weights to the sentences in the bag. These approaches aim to enhance the impact of accurate sentences while mitigating the influence of erroneous ones. Other ways involve using reinforcement learning or adversarial learning [8, 24, 26, 29] to select clean sentences from the bag to train the model. Nevertheless, recent research [4] indicates that bag-level DSRE methods have a limited effect on sentence-level prediction. Besides, bag-level methods are unable to assign a specific sentence label to each sentence in the bag and disregard the fact that all sentences in the bags are noisy samples.

Thus, sentence-level distantly supervised relation extraction has received increasing attention in recent years. Sentence-level DSRE methods typically employ sampling strategies to filter noisy data. Jia et al. [11] refined the DS dataset by identifying frequently occurring relation patterns. Ma et al. [18] utilized complementary labels to create negative samples and employed negative training to filter out noisy data. Li et al. [15] incorporated a small amount of external reliable data in their training process through meta-learning. Adjusting the loss function [5, 7] is also a commonly used method to reduce the impact of noisy samples. Despite the effectiveness of these methods in handling noisy data, they possess certain potential limitations, such as reliance on prior knowledge, subjectivity in sample construction, and access to external data. Different from previous work, our method iteratively executes noise divide and guided label generation to refine the DS-built data, independent of external data and prior knowledge.

3 Methodology

In this section, we introduce NDGR, a novel training framework for sentence-level DSRE. As shown in Fig. 1, our method comprises three main steps: (1) Divide noisy data from the DS-built dataset by modeling the loss distribution with a GMM (Sect. 3.1); (2) Generate high-quality pseudo-labels for unlabeled

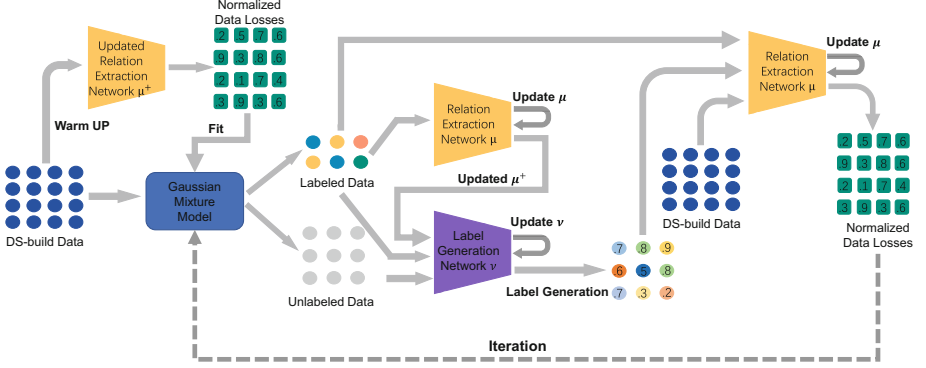


Fig. 1. An overview of the NDGR. There are three main steps: (1) Divide noisy data by employing a GMM to model the distribution of losses; (2) Generate high-quality pseudo labels for unlabeled data by guided label generation strategy; (3) Iterative training to further strengthen performance.

data using a guided label generation strategy (Sect. 3.2); (3) Iterative training based on (1) and (2) to further refine DS-built dataset (Sect. 3.3).

3.1 Noise Division by Loss Modeling

Previous research suggests that deep neural networks exhibit a swifter adaptation to clean data in contrast to noisy data [1], resulting the loss value incurred by clean data is lower than that of noisy data [2]. Hence, we attempt to separate noisy instances from the DS-built dataset by the loss value of each sample. Inspired by Li et al. [13], we aim to get the clean probability of each sample by fitting a Gaussian Mixture Model (GMM) [21] to the sample loss distribution.

Formally, we denote the input DS-built dataset as $\mathbf{D} = \{(S, Y) = \{(s_i, y_i)\}_{i=1}^N$, where $y_i \in \{1, \dots, C\}$ is the class label for the i^{th} input sentence s_i . In the initial stage of our method, we warm up the Relation Extraction Network (REN) on all DS-built data for a few epochs to get the initial loss distribution. Specifically, for a model with parameters θ , we have:

$$L(\theta) = \sum_{i=1}^N (l_i) = \sum_{i=1}^N \text{loss}(p_i, y_i) \quad (1)$$

$$p_i = M_\theta(s_i) \quad (2)$$

where p_i is the probability distribution of the relation. M_θ consists of the encoder module which converts the input sentence s_i into sentence representation h and the fully connected layer that applies h for classification.

However, we found the network would quickly overfit the noisy data during the warm-up phase, which resulted in most samples having similar normalized loss values close to zero (as shown in Fig. 2(a)). In this case, it's difficult for

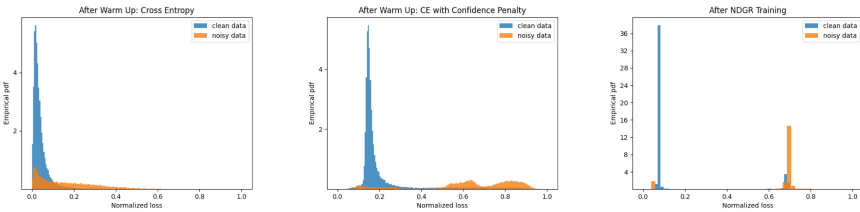
GMM to distinguish the clean and noisy samples based on the loss distribution. To address this issue, we penalize confident output distribution by adding a negative entropy $-H$ to cross-entropy loss during training [20].

$$H = -p_i \log(p_i) \quad (3)$$

The whole loss function as Eq. 4 shows:

$$L(\theta) = \sum_{i=1}^N \{loss(p_i, y_i) + p_i \log(p_i)\} \quad (4)$$

After maximizing the entropy, the normalized loss is distributed more evenly (as shown in Fig. 2(b)) and is easier to model by GMM. We then apply the Expectation-Maximum (EM) algorithm to fit a two-component Gaussian Mixture Model [21]. For each sample, we calculate a posterior probability $p(g|l_i)$ as its clean probability c_i , where g is the smaller mean Gaussian component. A threshold Th is established to partition the training data into two separate sets. The samples with c_i greater than Th are assigned to a clean labeled set X , otherwise are assigned to a noisy unlabeled set U .



(a) Standard CE

(b) CE with Penalty

(c) After NDGR Training

Fig. 2. Loss distribution when training on Noisy-TACRED. (a) Training using the standard cross-entropy loss function, may lead to overfitting and overly confident predictions. (b) Adding a negative entropy to cross-entropy loss leads to the normalized loss being distributed more evenly. (c) After NDGR training, the clean and noisy data are further separated.

3.2 Guided Label Generation and Exploitation

Following the isolation of noisy data from the dataset, the majority of preceding research employed exclusively clean data for training purposes, thereby overlooking the valuable information embedded in the noisy data. However, proper handling of noisy data can effectively improve the performance of the model. In this section, we introduce the guided label generation strategy, aimed at generating high-quality pseudo-labels for noisy data.

To avoid the gradual drift in the distribution of sentence features caused by the noise present in the generated pseudo labels [17], we construct two networks: the Relation Extraction Network (REN) which is trained to extract relations from unstructured text and the Label Generation Network (LGN) which has the same architecture as REN but is trained separately to generate pseudo labels for the unlabeled data.

To distinguish, we denote the parameters of REN as μ and the parameters of LGN as ν . Using the updated REN as a reference, we let the LGN learn to evaluate the quality of the generated labels. To optimize the ν , we adopt the loss function as follows:

$$L(\nu) = \sum_{i=1}^N (\text{loss}(M_{\mu^+}(s_i), y_i) + W\text{loss}(M_{\nu}(s_i), y_i)) \quad (5)$$

where $W \in [0, 1]$ is a manually set hyperparameter, μ^+ denotes the parameters of the REN after a gradient update based on the loss function defined in Eq. 4. In this way, we calculate the loss value and update ν using the updated parameters μ^+ . This can aid LGN in acquiring a deeper understanding of the training procedure employed by REN. To prevent the accumulation of errors caused by noise in the generated labels during training, LGN is trained on labeled set X .

After optimizing the LGN for a few epochs, we generate pseudo labels for the unlabeled set U . For each sample u_i in U , the generated pseudo label is defined:

$$\text{label} = \arg \max(M_{\nu^+}(u_i)) \quad (6)$$

where ν^+ is the updated parameters of the LGN.

The relation associated with the highest probability after softmax is considered as the pseudo-label. We utilize these generated pseudo-labels by amalgamating the re-labeled set R with the labeled set X to form an enhanced dataset. Subsequently, this dataset is employed to retrain REN, leading to performance improvement.

3.3 Iterative Training Algorithm

While dividing noisy data by modeling the loss distribution with a GMM and implementing the guided label generation strategy for re-labeling the noisy data can refine the dataset and enhance performance, it fails to fully exploit the potential of each component. Therefore, we employ iterative training to further enhance performance.

As shown in Fig. 1, for each iteration, we firstly divide the DS-built into labeled data and unlabeled data by modeling the data loss distribution with GMM (before the first iteration, we warm up the model to get the initial normalized loss distribution). Following M epochs of REN training using labeled data, we leverage the updated REN as a reference to optimize LGN. Subsequently, the updated LGN is employed for generating pseudo-labels pertaining to the unlabeled data. By amalgamating the labeled and re-labeled data, a novel

refined dataset is formulated. Prior to retraining REN with the refined dataset, a model re-initialization step is executed to avert overfitting. This process ensures that the models are optimized using a high-quality dataset and incorporates randomness, thereby improving the robustness of our method. Finally, we input the origin DS-build data into the updated REN to obtain the new normalized loss distribution for fitting the GMM and perform re-initialization of both the REN and LGN to enter the next iteration. Figure 2(c) shows the loss distribution after NDGR training. As seen, there is a substantial margin in the loss values between most of the noisy data and clean data. Most of the noisy data has been successfully separated, with only an acceptable amount of clean data being misclassified, demonstrating the robust de-noising ability of NDGR.

4 Experiments and Analysis

To evaluate the efficacy of the proposed method, we divided the experiment into two parts and conducted tests on two datasets: (1) The first part is to validate the effectiveness of our proposed method at the sentence-level evaluation. Numerous previous DSRE approaches employ a held-out evaluation, where both the training and test sets are constructed using the DS method. According to the study, Gao et al. [6] suggest that using a held-out test set cannot accurately demonstrate the model’s performance. Hence, we utilized a manually-labeled test set to evaluate the model. (2) In the second part, a series of experiments are designed to evaluate the efficacy of each component in NDGR. Since the DS-build dataset cannot label whether this instance is mislabeled, we construct a noisy dataset called Noisy-TACRED from the manually labeled dataset.

4.1 Datasets

We evaluate our method on two widely-used datasets: the NYT dataset and the Noisy-TACRED dataset.

NYT: Riedel et al. [25] constructed this dataset by aligning the New York Times corpus with entity-relationship triples from Freebase. The original training and test sets are both established using the DS method, encompassing noisy data. For a more precise evaluation, we employ the original training set alongside a manually annotated sentence-level test set [11].

Noisy-TACRED: The original TACRED dataset, constructed by Zhang et al. [33], comprises 80% of instances labeled as “NA”. The “NA” rate is similar to the NYT dataset which is constructed by the DS method, hence analysis on this dataset is more reliable. To create the Noisy-TACRED dataset, we select noisy instances randomly with a noisy ratio of 30%. For each noisy instance, a noisy label is assigned by randomly selecting a label from a complementary class. The selection probability of a label is determined by its class frequency, this approach helps to preserve the original distribution of data.

4.2 Baseline Models

We compare our method with multiple strong baselines as follows:

PCNN [29]: A *bag-level* method with multi-instance learning to address the wrong label problem.

PCNN+SelATT [16]: A *bag-level* de-noise method uses the attention mechanisms to reduce the impact of noisy data.

PCNN+RA_BAG_ATT [27]: A *bag-level* method that utilizes the inter-bag and intra-bag attention mechanisms to alleviate noisy instances.

CNN+RL_1 [24]: A *bag-level* method that applies reinforcement learning to recognize the false positive samples and then the filtered data reallocated as negative samples.

CNN+RL_2 [4]: A *sentence-level* method which incorporates reinforcement learning to jointly train a RE model for relation extraction and a selector to filter the potential noisy samples.

ARNOR [11]: A *sentence-level* method that selects reliable instances by rewarding high attention scores on specific patterns.

SENT(BiLSTM) [18]: A *sentence-level* DSRE method filters the noisy data by negative training and performs a re-label process to transform the noisy data into useful data.

CNN [30], **BiLSTM** [32] and **BERT** [3] are widely-used models for *sentence-level* relation extraction without denoising method.

4.3 Implementation Details

Our proposed method employs the BiLSTM as the sentence encoder. 50-dimensional GloVe vectors [16] are used as word embeddings during training. Furthermore, we incorporate 50-dimensional randomly initialized position and entity type embeddings throughout all training phases. The hidden size of the BiLSTM is set to 256. It is optimized using the Adam optimizer with a learning rate of $1e-5$. The weight parameter W in Eq. 4 is assigned a value of $5e-1$.

The hyperparameters are tuned by performing a grid search on the validation set. When training on the NYT dataset, we first warm up the REN using all DS-built data for 4 epochs. We then perform a total of 8 iterations, with each iteration involving training the model for 15 epochs. The data-divide threshold is set to $Th = 0.7$. During the training phase on the Noisy-TACRED, the REN is initially warmed up using all DS-built data for 45 epochs. Following this warm-up phase, the model goes through 8 iterations, where each iteration involves 50 epochs of training. The training procedure utilizes a data-divide threshold Th of 0.6 and sets the learning rate to $5e-4$.

4.4 Sentence-Level Evaluation

Table 1 shows the results of our method and other baseline models on sentence-level evaluation. Consistent with previous methods [11, 15], we calculate Micro-Precision (Prec.), Micro-Recall (Rec.), and Micro-F1 (F1) to evaluate the effectiveness of our approach. According to the results, we can observe that: (1)

Table 1. Results of our method and other baseline models on sentence-level evaluation. The first part of the table is ordinary RE methods without denoising and the second part of the table is distant RE methods. The results with “*” are bag-level methods and the results with “†” are sentence-level methods.

Method	Dev			Test		
	Prec.	Rec.	F1	Prec.	Rec.	F1
CNN†	38.32	65.22	48.28	37.75	64.54	46.01
PCNN*	36.09	63.66	46.07	36.06	64.86	46.35
BiLSTM†	36.71	66.46	47.29	35.52	67.41	46.53
BERT†	34.78	65.17	45.35	36.19	70.44	47.81
PCNN+SelATT*	46.01	30.43	36.64	45.51	30.03	36.15
PCNN+RA_BAG_ATT*	49.84	46.90	48.33	56.76	50.60	53.50
CNN+RL_1*	37.71	52.66	43.95	39.41	61.61	48.07
CNN+RL_2†	40.00	59.17	47.73	40.23	63.78	49.34
ARNOR†	62.45	58.51	60.36	65.23	56.79	60.90
SENT (BiLSTM)†	66.71	57.27	61.63	71.22	59.75	64.99
NDGR (BiLSTM)	74.34	58.46	65.45	78.30	56.97	65.95

When trained with DS-built data without de-noise, all baseline models performed poorly. Even the highly acclaimed pre-trained language model BERT, renowned for its superior performance in sentence-level relation extraction tasks on clean datasets, demonstrated subpar results. This phenomenon underscores the substantial impact of noisy samples in the dataset on model training, particularly for pre-trained language models, which are prone to overfitting such data. (2) The bag-level methods demonstrate poor performance in sentence-level evaluation, indicating their unsuitability for downstream tasks that require precise sentence labels. Therefore, it is imperative to explore sentence-level methods for DSRE. (3) The proposed NDGR method achieves a significant improvement over previous sentence-level de-noise methods. Our implementation utilizing BiLSTM as the sentence encoder results in a 0.96% improvement in the F1 score compared to SENT. Additionally, it exhibits significantly higher precision while maintaining comparable recall. These outcomes highlight the effectiveness of our strategy for data division and guided label generation.

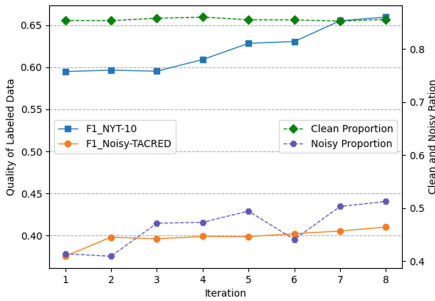
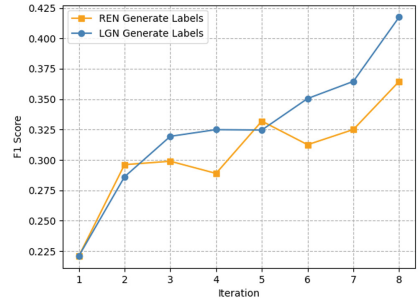
4.5 Analysing on Noisy-TACRED

In this section, we analyze the efficacy of noise division and label generation process on the Noisy-TACRED dataset.

Table 2. Model performance on clean-TACRED and noisy-TACRED.

	Method	Prec.	Rec.	F1
Clean TACRED	BiLSTM+ATT	67.7	63.2	65.4
	BiLSTM	61.4	61.7	61.5
Noisy TACRED	BiLSTM+ATT	32.8	43.8	37.5
	BiLSTM	37.8	45.5	41.3
	NDGR (BiLSTM)	86.4	43.3	57.7

Evaluation on Noisy-TACRED: We trained in Clean-TACRED and Noisy-TACRED respectively, and the results are shown in Table 2. Comparing the results on two datasets, we can find the performance of baseline models degraded significantly, the F1 value of BiLSTM+ATT decreased by 27.9, while the F1 value of BiLSTM decreased by 20.2. By employing our proposed NDGR on the noisy data, the BiLSTM model demonstrates comparable performance to the model trained on clean data. This finding indicates that our methodology effectively mitigates the impact of incorrectly labeled data in the DS-built dataset.

**Fig. 3.** Experimental Details of Divide Data by Loss Modeling**Fig. 4.** The quality of the pseudo labels generated by REN and LGN.

Effects of Divide Data by Loss Modeling: As described in Sect. 3.1, we employ GMM to model the loss distribution, and subsequently leverage the posterior probability to partition the noisy data within the dataset constructed by DS. To demonstrate the efficacy of the loss modeling by GMM in distinguishing between clean and noisy data. We first evaluated the quality of the labeled data, we trained the model only with the labeled data and ignored the unlabeled data on both NYT and Noisy-TACRED. Additionally, we compute the ratio of clean data within the labeled set and noisy data within the unlabeled set on Noisy-TACRED.

The results are shown in Fig. 3, we can observe: 1) The F1 score exhibits a gradual increase with each iteration for both the NYT and Noisy-TACRED datasets, suggesting an improvement in the quality of the labeled data. 2) As the iteration progressed, the ratio of clean data in the labeled set and the ratio of noise data in the unlabeled set both increased. The proportion of clean data can reach approximately 85%, while the proportion of noisy data amounts to around 51%. These observations affirm the effective noise filtering capability of GMM-based loss modeling within the dataset constructed using DS.

Effects of Guided Pseudo Label Generation: As described in Sect. 3.2, once the noisy data has been separated from the training dataset, we employ the guided label generation strategy for converting the unlabeled data into valuable training data. To verify the effectiveness of this strategy, we separately used REN and LGN to generate pseudo labels and calculated the F1 score between the generated labels and the original labels of the TACRED, the results are shown in Fig. 4. As seen, the pseudo-labels generated directly by REN exhibit inferior performance compared to those generated by LGN. This demonstrates that the label generation strategy we employ can reduce noise in generating labels and improve performance.

Table 3. Ablation study on NYT dataset

Components	Prec.	Rec.	F1
NDGR	78.30	56.97	65.95
w/o Guided Label Generation	58.31	64.09	61.06
w/o Re-initialization	55.11	63.47	58.99
w/o Noise Division	45.11	69.97	54.85
w/o Confidence Penalty	47.64	62.54	54.08

4.6 Ablation Study

We conduct an ablation study to demonstrate the contribution of each component in our proposed method on the NYT dataset. We specifically assess the performance by removing certain components, including guided label generation, re-initialization, noise divide, and confidence penalty. The results are shown in Table 3, we can observe that: 1) Without the guided label generation strategy, instead of employing LGN for label generation, we employ REN to directly generate labels. The presence of noise in the generated labels results in error accumulation during iterations, causing a progressive drift in the distribution of sentence features, thereby impacting the overall system performance. 2) Re-initialization has a great contribution to the performance. In the label generation phase, despite utilizing labeled data for training REN and LGN with the intention of mitigating the impact of noise, it is inevitable that certain noise data will

become incorporated and the models will adapt to them. With re-initialization, REN will initialize the overfitting parameters and retrain on the refined dataset, thus contributing to better performance. 3) Noise division significantly impacts the performance. Without noise division, we use the original DS-built data to optimize the REN and LGN, then generate pseudo-labels for all DS-built data. However, due to the presence of noise during the training process, the quality of these labels diminishes, resulting in inferior performance. Moreover, as the iterations progress, the performance further deteriorates. 4) Confidence penalty contributes a lot to the performance. In the absence of the confidence penalty, most of the normalized loss values are comparable and tend toward zero. This makes it challenging for the GMM to effectively filter out noisy samples based on the distribution of losses. The presence of numerous mislabeled data has significantly affected the subsequent label generation process, resulting in a decline in the quality of the newly constructed training dataset.

5 Conclusion

In this paper, we propose NDGR, a novel sentence-level DSRE training framework that incorporates noise division, guided label generation, and iterative training. Specifically, NDGR first separates noisy data from the training dataset by modeling the data loss distribution with a GMM. Next, we assign pseudo-labels for unlabeled data using a guided label generation strategy to reduce the noise in the generated pseudo-labels. Through iterative execution of noise division and guided label generation, NDGR helps re-fine the noisy DS-built data and enhance the performance. Extensive experiments on widely-used benchmarks have demonstrated that our method has significant improvement in sentence-level relation extraction and de-noise effect.

Acknowledgements. This work was supported by Sichuan Science and Technology Program (2019ZDZX0006, 2020YFQ0056), Science and Technology Service Network Initiative (KFJ-STS-QYZD-2021-21-001), and the Talents by Sichuan provincial Party Committee Organization Department.

References

1. Arpit, D., et al.: A closer look at memorization in deep networks. In: International Conference on Machine Learning, pp. 233–242. PMLR (2017)
2. Chen, P., Liao, B.B., Chen, G., Zhang, S.: Understanding and utilizing deep neural networks trained with noisy labels. In: International Conference on Machine Learning, pp. 1062–1070. PMLR (2019)
3. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Minneapolis, Minnesota, June 2019, pp. 4171–4186. Association for Computational Linguistics (2019). <https://doi.org/10.18653/v1/N19-1423>. <https://aclanthology.org/N19-1423>

4. Feng, J., Huang, M., Li, Z., Yang, Y., Zhu, X.: Reinforcement learning for relation classification from noisy data (2018)
5. Fu, B., Peng, Y., Qin, X.: Learning with noisy labels via logit adjustment based on gradient prior method. *Appl. Intell.* **53**, 24393–24406 (2023). <https://doi.org/10.1007/s10489-023-04609-1>
6. Gao, T., et al.: Manual evaluation matters: reviewing test protocols of distantly supervised relation extraction. arXiv preprint [arXiv:2105.09543](https://arxiv.org/abs/2105.09543) (2021)
7. Ghosh, A., Kumar, H., Sastry, P.S.: Robust loss functions under label noise for deep neural networks. In: *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 31 (2017)
8. Han, X., Liu, Z., Sun, M.: Denoising distant supervision for relation extraction via instance-level adversarial training. arXiv preprint [arXiv:1805.10959](https://arxiv.org/abs/1805.10959) (2018)
9. Han, X., Liu, Z., Sun, M.: Neural knowledge acquisition via mutual attention between knowledge graph and text. In: *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32 (2018)
10. Hoffmann, R., Zhang, C., Ling, X., Zettlemoyer, L., Weld, D.S.: Knowledge-based weak supervision for information extraction of overlapping relations. In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Portland, Oregon, USA, June 2011*, pp. 541–550. Association for Computational Linguistics (2011). <https://aclanthology.org/P11-1055>
11. Jia, W., Dai, D., Xiao, X., Wu, H.: ARNOR: attention regularization based noise reduction for distant supervision relation classification. In: *Meeting of the Association for Computational Linguistics* (2019)
12. Jiang, X., Wang, Q., Li, P., Wang, B.: Relation extraction with multi-instance multi-label convolutional neural networks. In: *Proceedings of the 26th International Conference on Computational Linguistics: Technical Papers, COLING 2016*, pp. 1471–1480 (2016)
13. Li, J., Socher, R., Hoi, S.C.: DivideMix: learning with noisy labels as semi-supervised learning. arXiv preprint [arXiv:2002.07394](https://arxiv.org/abs/2002.07394) (2020)
14. Li, Y., et al.: Self-attention enhanced selective gate with entity-aware embedding for distantly supervised relation extraction. In: *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, pp. 8269–8276 (2020)
15. Li, Z., et al.: Meta-learning for neural relation classification with distant supervision. In: *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pp. 815–824 (2020)
16. Lin, Y., Shen, S., Liu, Z., Luan, H., Sun, M.: Neural relation extraction with selective attention over instances. In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2124–2133 (2016)
17. Liu, S., Davison, A., Johns, E.: Self-supervised generalisation with meta auxiliary learning. In: *Advances in Neural Information Processing Systems*, vol. 32 (2019)
18. Ma, R., Gui, T., Li, L., Zhang, Q., Zhou, Y., Huang, X.: SENT: sentence-level distant relation extraction via negative training. arXiv preprint [arXiv:2106.11566](https://arxiv.org/abs/2106.11566) (2021)
19. Mintz, M., Bills, S., Snow, R., Jurafsky, D.: Distant supervision for relation extraction without labeled data. In: *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pp. 1003–1011 (2009)

20. Pereyra, G., Tucker, G., Chorowski, J., Kaiser, Ł., Hinton, G.: Regularizing neural networks by penalizing confident output distributions. arXiv preprint [arXiv:1701.06548](https://arxiv.org/abs/1701.06548) (2017)
21. Permuter, H., Francos, J., Jermyn, I.: A study of Gaussian mixture models of color and texture features for image classification and segmentation. *Pattern Recogn.* **39**(4), 695–706 (2006)
22. Qin, P., Xu, W., Wang, W.Y.: Robust distant supervision relation extraction via deep reinforcement learning. In: Meeting of the Association for Computational Linguistics (2018)
23. Qin, P., Xu, W., Wang, W.Y.: DSGAN: generative adversarial training for distant supervision relation extraction. arXiv preprint [arXiv:1805.09929](https://arxiv.org/abs/1805.09929) (2018)
24. Qin, P., Xu, W., Wang, W.Y.: Robust distant supervision relation extraction via deep reinforcement learning. arXiv preprint [arXiv:1805.09927](https://arxiv.org/abs/1805.09927) (2018)
25. Riedel, S., Yao, L., McCallum, A.: Modeling relations and their mentions without labeled text. In: Balcázar, J.L., Bonchi, F., Gionis, A., Sebag, M. (eds.) ECML PKDD 2010. LNCS (LNAI), vol. 6323, pp. 148–163. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-15939-8_10
26. Shang, Y., Huang, H.Y., Mao, X.L., Sun, X., Wei, W.: Are noisy sentences useless for distant supervised relation extraction? In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 8799–8806 (2020)
27. Ye, Z.X., Ling, Z.H.: Distant supervision relation extraction with intra-bag and inter-bag attentions. arXiv preprint [arXiv:1904.00143](https://arxiv.org/abs/1904.00143) (2019)
28. Yuan, Y., et al.: Cross-relation cross-bag attention for distantly-supervised relation extraction. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 419–426 (2019)
29. Zeng, D., Liu, K., Chen, Y., Zhao, J.: Distant supervision for relation extraction via piecewise convolutional neural networks. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pp. 1753–1762 (2015)
30. Zeng, D., Liu, K., Lai, S., Zhou, G., Zhao, J.: Relation classification via convolutional deep neural network. In: Proceedings of the 25th International Conference on Computational Linguistics: Technical Papers, COLING 2014, pp. 2335–2344 (2014)
31. Zeng, X., He, S., Liu, K., Zhao, J.: Large scaled relation extraction with reinforcement learning. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32 (2018)
32. Zhang, S., Zheng, D., Hu, X., Yang, M.: Bidirectional long short-term memory networks for relation classification. In: Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation, pp. 73–78 (2015)
33. Zhang, Y., Zhong, V., Chen, D., Angeli, G., Manning, C.D.: Position-aware attention and supervised data improve slot filling. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, pp. 35–45 (2017). <https://nlp.stanford.edu/pubs/zhang2017taced.pdf>